**Emotion rating online validation study**

Prior to study 1, we ran an online stimulus validation study to ensure that the faces would receive comparable ratings in perceived emotionality. At the time of this online validation study (June 2018), we had not yet added the third control condition (pareidolic faces), so only the emotional content of unique human and robot faces were rated. Furthermore, as this was the first set of stimuli, there were less unique images (12 images per condition) compared to study 2 (24 images per condition). The validation experiment was presented in Jisc Online Surveys (formerly Bristol Online Surveys). Participants rated 18 unique robot and 18 unique human faces (male and female) on a bespoke semantic differential scale between ‘1 – sad’ to ‘7 – happy’. ‘4’ was considered ‘neutral’, for the purpose of the analysis. The scale was made for this study. 84 participants (age: *M=*34.67, *SD*=11.77) completed the rating study. Most participants were female (*n*=64) and most reported never having interacted with a robot before (*n*=61). The participants were recruited via advertisements on social media.

Figure S1. The three removed images are labelled (robot 7, 11, and 12) – as it was costly to exclude images, robot 2 (on the border of > 1 SD below the mean was kept in the set. The bold dots represent the mean rating scores for each image, and the bars represent the standard error.

While the two groups didn’t differ in mean ratings at first glance (human faces: *M*= 3.69, *SD*=1.08, robot faces: *M*=3.97, *SD*=1.39), ordinal logistic regression with the ‘ordinal’ package (Christensen, 2019) suggests that human faces were rated more negatively (estimate = -.37, SE= .07, p<.001). Following this result, we inspected the mean ratings of the individual stimuli visually and discovered that the robotic faces were rated much more variably than the human faces. As the stimulus exclusions were costly (i.e. the time and effort to replace and re-process all images), and we had to work towards keeping at least 12 unique images within the pool of 18 images in each condition, we removed the strongest outliers in the robot condition (robots 2, 6, 7, 10, 11, & 12) and removed those human faces that were rated more negatively than the average, with 3 male and 3 female faces each (19, 22, 23, 28, 33 & 35). While the procedure we followed is not optimal (limited by the time and stimulus availability constraints), we gained valuable insights. The robot faces we considered “neutral” upon selection were, in fact, perceived not as unambiguously neutral by the raters, and despite selecting human faces from the neutral condition of the Radboud Faces Database (Langner et al., 2010), they were perceived slightly more negatively than the midpoint of our scale.